General Biostatistics

Part 9

Common Statistical Methods

Simple Linear Regression - Class 5A

Marie Diener-West, Ph.D.
Department of Biostatistics
Johns Hopkins University
Bloomberg School of Public Health

Outline

- Association
- Correlation
- Straight line relationship
- Simple linear regression
- Method of Least Squares
- Interpretation of least squares coefficients
- Example

Association

- Express the relationship or association between two variables
- Can be measured in different ways depending on the nature of the variables
  - continuous (e.g. height and weight)
  - ordinal (e.g. Apgar score and birth weight category)
  - nominal (e.g. vital status and cancer treatment)

Measures of Association

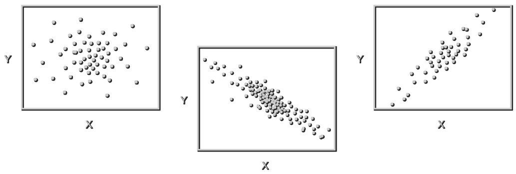
- Chi-squared statistic
  - Association between 2 nominal variables
- Pearson correlation coefficient
  - Linear relationship
- Spearman rank correlation coefficient
- Kappa statistic
  - Agreement between raters

Correlation Analysis

- Describe the straight line relationship between two continuous variables
  - Correlation analysis
    - Measuring strength and direction of relationship
  - Regression analysis
    - Predicting or estimating the value of one variable based on the value of the other variable

Correlation Analysis

- Visually inspect the data
  - no, negative, positive correlation?

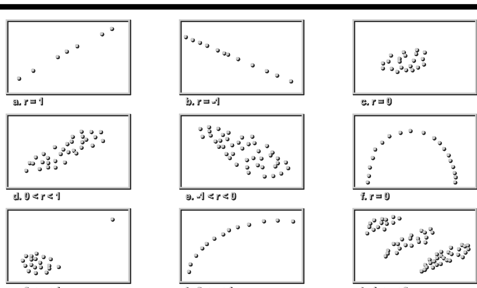


Correlation Analysis

- Sample correlation coefficient, r
  - Independent of the units used to measure the variables
    - r = +1, perfect positive association
    - r = 0, no association
    - r = -1, perfect negative association
- Guidelines for interpretation (+ or -)

0 to 0.25	little or no relationship
0.25 to 0.50	fair
0.50 to 0.75	moderate to good
0.75 to 1.00	very good to excellent

Examples of Correlation



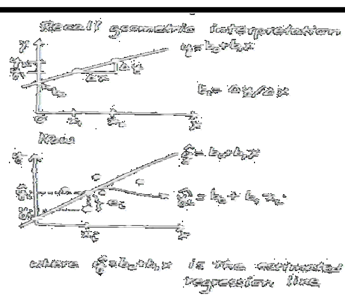
Correlation Analysis

- Test Ho: ρ =0 where ρ is the true population correlation coefficient
- Construct confidence interval for ρ

Simple Linear Regression (SLR)

- A linear regression describes a response measure, Y, the <u>dependent variable</u>, as a function of an explanatory variable, X, the <u>independent variable</u>
- Goal: predict or estimate the value of Y based on the value of X

Straight Line Relationship

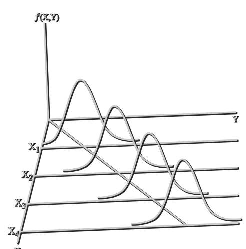


Method of Least Squares

- $Y = \beta_0 + \beta_1 X + \varepsilon$
- β_0 = the y-intercept
- β_1 = the slope = $\Delta Y / \Delta X$

Assumptions of SLR

- L - linear relationship
- I - independent Y's
- N - normally distributed
- E - equal variance



Method of Least Squares

- The "best" line is determined by finding the estimates of β_0 and β_1 which minimize the sum of the squared differences between an observed point and a fitted point on the line (e.g. minimizes the sum of squared "error" terms)

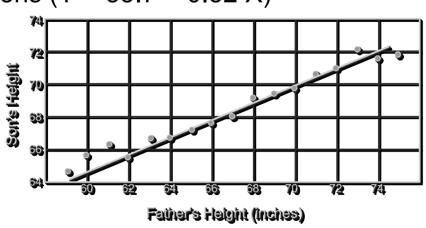
Least Squares Estimators

$$b_0 = \overline{Y} - b_1 \overline{X}$$

$$b_1 = \frac{\sum_i X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_i X_i^2 - n\overline{X}^2}$$

Galton’s Linear Regression

- Galton’s study of heights of fathers and sons (Y = 33.7 + 0.52 X)

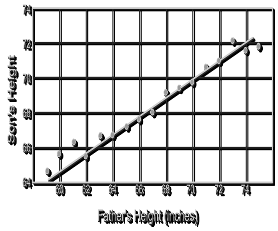


Galton’s Linear Regression

- How do we interpret Y = 33.7 + 0.52 X?
- $Y = b_0 + b_1 X$
- Y = son’s height; X = father’s height
- b_0 = expected son’s height when father’s height is 0 inches
- b_1 = the difference in expected heights of sons whose fathers’ heights differ by one inch

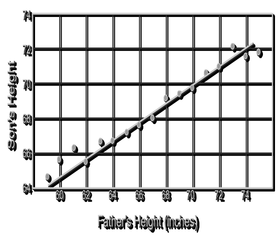
Galton’s Linear Regression

- $Y = b_0 + b_1 X = 33.7 + 0.52X$
- What is b_0 ?
- b_0 = 33.7 inches = expected son’s height when father’s height is 0 inches



Galton’s Linear Regression

- What is b_1 ?
- b_1 = 0.52 inch difference in expected heights of sons whose fathers’ heights differ by one inch
- “Reversion” or “regression to the mean”



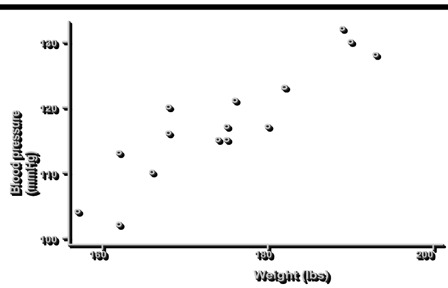
Example

Weight (lbs)	Blood Pressure (mm Hg)
157	104
162	113
162	102
168	120
166	110
168	116
174	115
175	115
176	121
175	117
180	117
182	123
193	128
189	132
190	130

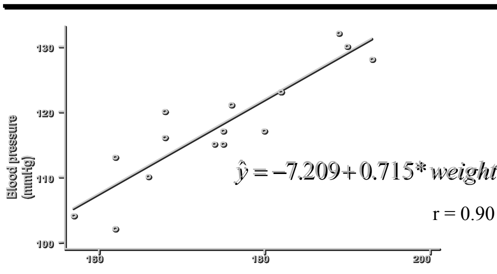
Example

- The mean blood pressure for this sample of 15 subjects is 117.5 mm Hg
- The sample variance is $s^2 = 74.3$ (mm Hg)^2
- The sample standard deviation is $s = 8.6$ mm Hg
- How much of the variability in blood pressure is due to its relationship with weight?

Example



Example



Example

$$\hat{y} = b_0 + b_1 * \text{weight}$$

$$\hat{y} = -7.209 + 0.715 * \text{weight}$$

- Simple linear regression line
- How do we interpret this line?

Example

$$\hat{y} = -7.209 + 0.715 * \text{weight}$$

- Based on a simple linear regression analysis of blood pressure on weight among these 15 subjects, the estimated change in the expected value of blood pressure is 0.715 mm Hg per each one pound change in weight (95%CI: 0.511, 0.919)

Example

$$\hat{y} = -7.209 + 0.715 * \text{weight}$$

- Based on this estimated regression line, we could predict blood pressure for a person weighing 175 pounds:

$$\hat{y} = -7.209 + 0.715(175) = 118 \text{ mm Hg}$$

Example

- The correlation coefficient r = 0.9
- r^2 = 0.81 which indicates that 81% of the variability in blood pressure can be explained by its linear relationship with weight
- From the linear regression, a better estimate of σ^2 = 14.8   (σ = 3.8)

ANOVA for Regression

- Partition the total variation in the Y's into two components: the variation *explained* by the regression (the linear relationship between Y and X) and the variation *not explained* by the regression (error)
- SST= SSR + SSE where
  - SST= total sum of squares
  - SSR= regression sum of squares
  - SSE = error sum of squares

ANOVA for Regression

```
0.  regress bp wt

  Source |       SS       df       MS                  Number of obs =      15
---------+------------------------------               F(  1,    13) =   57.30
   Model |  847.45661      1  847.45661                Prob > F      =  0.0000
Residual |  192.276723    13  14.7905172               R-squared     =  0.8151
---------+------------------------------               Adj R-squared =  0.8008
   Total |  1039.73333    14  74.2666667               Root MSE      =  3.8458

------------------------------------------------------------------------------
      bp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      wt |   .7149924    .094457     7.569   0.000       .5109303    .9190544
   _cons |     -7.209    16.5095    -0.437   0.670      -42.8756    28.45759
------------------------------------------------------------------------------
```

Summary

- Simple linear regression analysis explores the linear relationship between one independent (predictor) variable and one dependent (response) variable.
  - Allows assessment of association (correlation)
  - Provides prediction
- Linear regression analysis can be extended to multiple predictor variables.